Galaxy for Data Provenance

Tom Doak
Le-Shin Wu
Carrie Ganote
National Center for Genome Analysis Support

July 16, 2014

INDIANA UNIVERSITY

Galaxy for Data Provenance

So...

What does data provenance mean?

The ability to reproduce and validate the scientific methods and outcomes given sufficient documentation of the data

Why is this important?

WOW, this unicorn mantis has never been seen by mankind!

I better get my camera…

Why is this important?

Without sufficient evidence to be able to reproduce your research, the conclusions are less compelling

NOOOOOOOOOOOOOOO
My dissertation!!!!

What needs to be considered?

- Raw data – how was it generated
- What software was used at each step (version, environment, operating system)
- What parameters were used
- How were intermediate files handled (copied, renamed, deleted)
- Backtracking, param sweeps, etc

How is provenance handled?

Common practices for analysis tracking involve:

- Lab Notebooks – paper and pen
- Naming files and directories descriptively
- Readme type files with each step documented
- Wiki pages for tracking steps
- Workflow-specific software
- Complicated custom solutions involving hashing everything

Pitfalls of these approaches

- Loss and damage potential
- They depend on the detail of the researcher
- Can involve more work than the actual work
- They may fail to make sense when revisited a year+ later

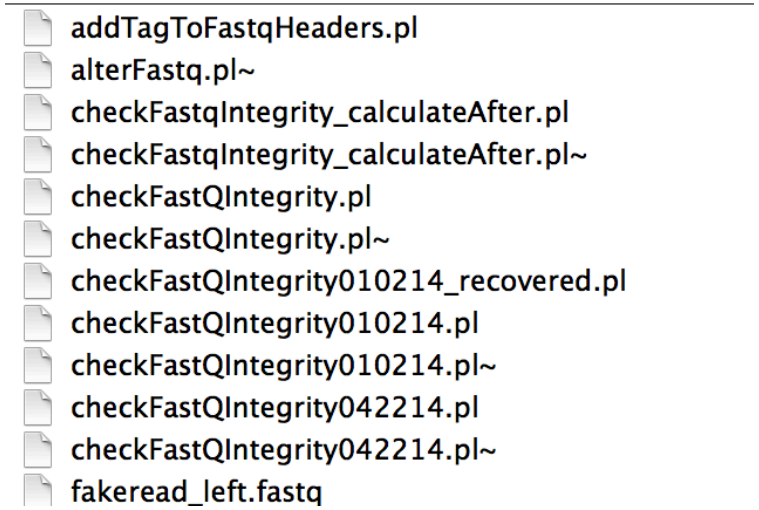
Does this look familiar?

Name	Date Modified	Date Created
visuals	3/12/13, 2:43 PM	Feb 11, 2013
TripleChecking	1/30/13, 3:30 PM	Jan 30, 2013
TrackedOnline	1/25/13, 5:37 PM	Oct 26, 2012
Temperature	4/30/13, 3:48 PM	Nov 19, 2012
Raw_Inputs	6/22/14, 7:44 PM	Apr 30, 2013
R-code-fo…p-covars.r	2/14/13, 5:12 PM	Feb 14, 2013
partek	2/21/13, 5:34 PM	Feb 13, 2013
Papers	7/8/14, 11:49 AM	Apr 30, 2013
Oct2013	11/12/13, 11:21 PM	Oct 31, 2013
Learnings	4/9/13, 3:02 PM	Nov 16, 2012
Deadends	5/1/13, 4:59 PM	Jan 25, 2013
August2013	8/21/13, 3:03 PM	Aug 21, 2013
Analysis	6/23/14, 11:02 AM	Dec 4, 2012,

Does this look familiar?

Adequate documentation practices are difficult to teach and exercise

addTagToFastqHeaders.pl_042414.11:33:28.temp
addTagToFastqHeaders.pl_042514.16:02:37.temp
addTagToFastqHeaders.pl_042514.16:24:18.temp
addTagToFastqHeaders.pl042814_042814.16:39:27.temp
addTagToFastqHeadersSkipmismatches.pl_050714.17:02:27.temp
addTagToFastqHeadersSkipmismatches.pl_050714.17:07:05.temp
addTagToFastqHeadersSkipmismatches.pl_050714.17:08:27.temp
addTagToFastqHeadersSkipmismatches.pl_050714.17:09:58.temp
addTagToFastqHeadersSkipmismatches.pl_050714.17:13:38.temp
addTagToFastqHeadersSkipmismatches.pl_050714.17:17:25.temp



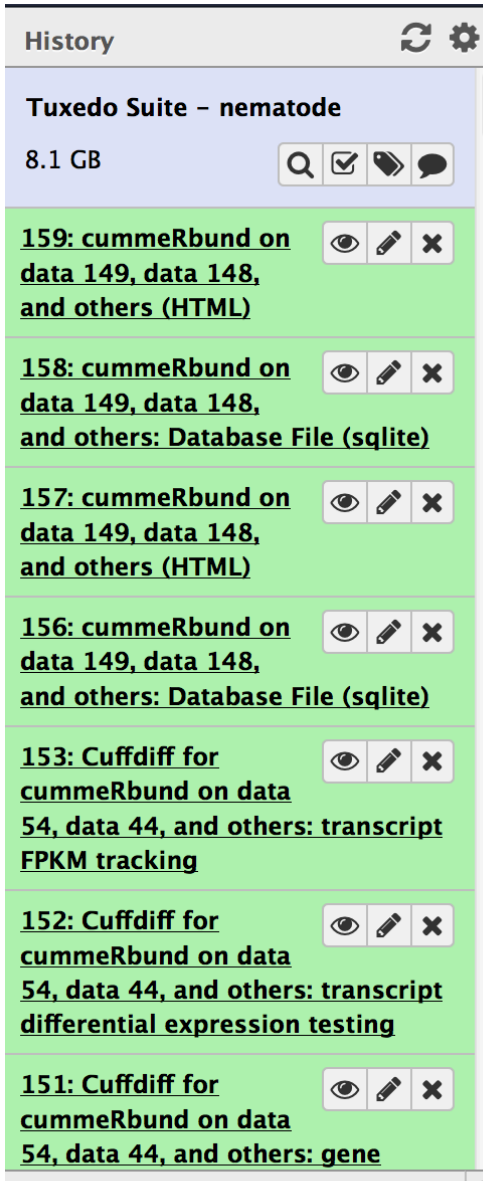
Version Control?

We can't just use Github

Great for tracking changes in a code base, not great for tracking a genomics project and parameter sweeping

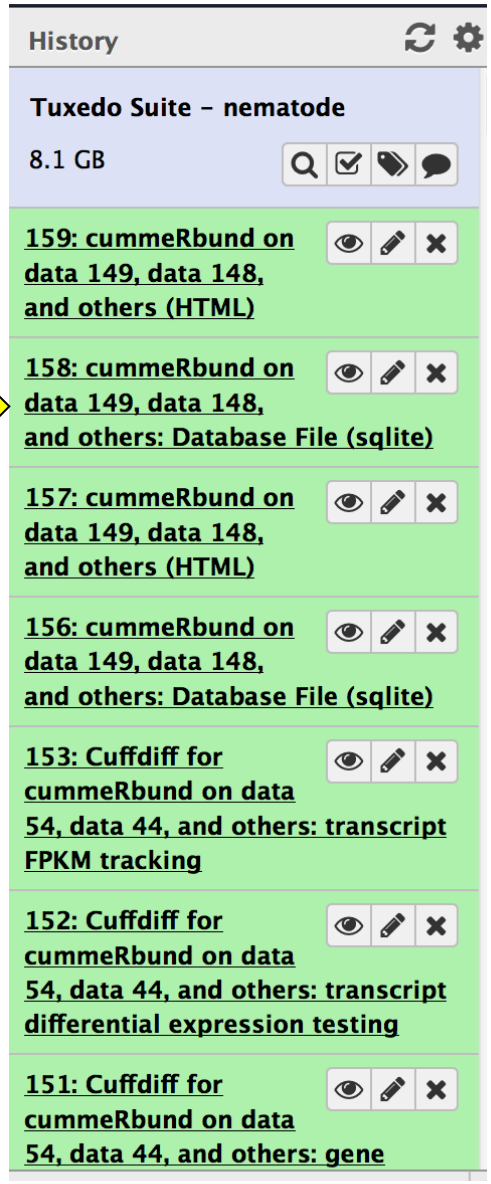
How does Galaxy help?

- History tracks every change made to a file
- When copying files to a new history/ sharing files with other users, hints about their origin are generated
- Galaxy uses version control to keep track of the exact tool wrapper used
- New features in Galaxy allow users to specify which version of the software the tool uses



Shortcomings of Galaxy

- Generated history names are hard to parse
- Under normal settings, user may not know for sure what version of (software, database, etc) they have used
- Does not prevent human error – renaming wrong file, admin error, etc.

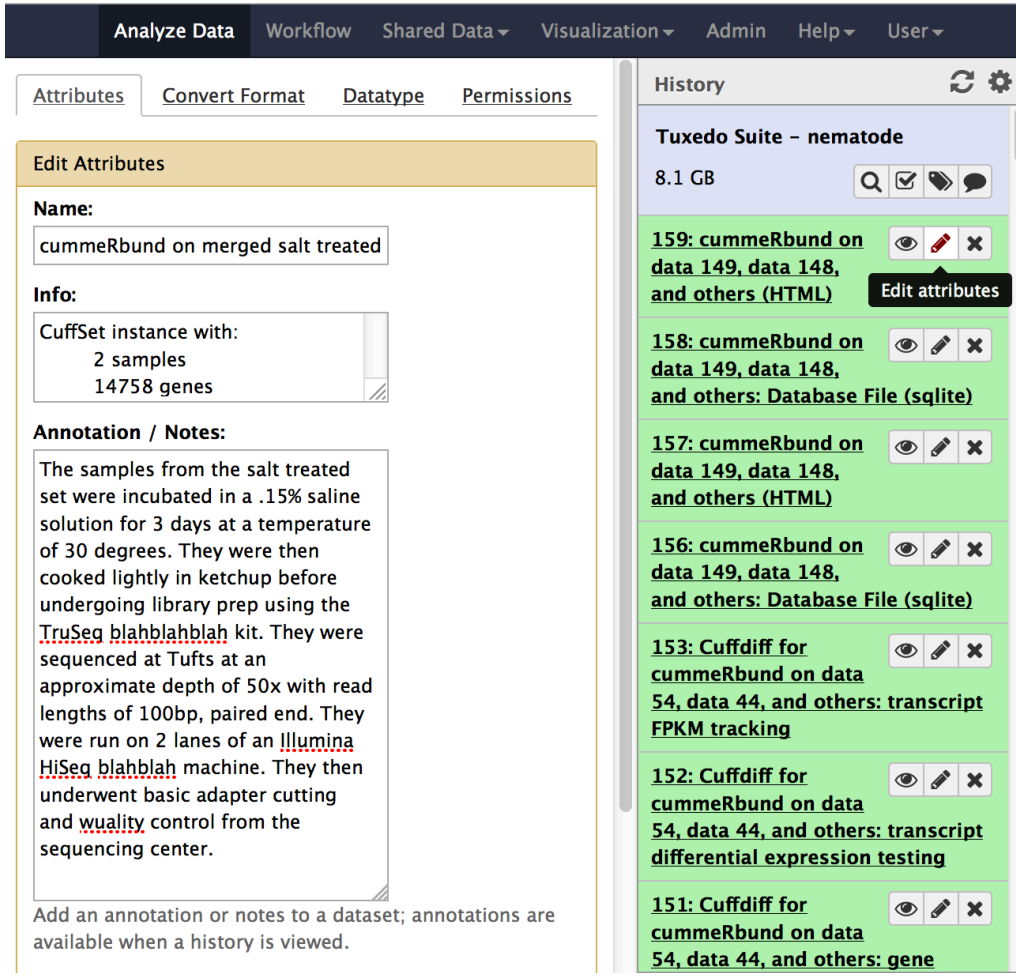


What?

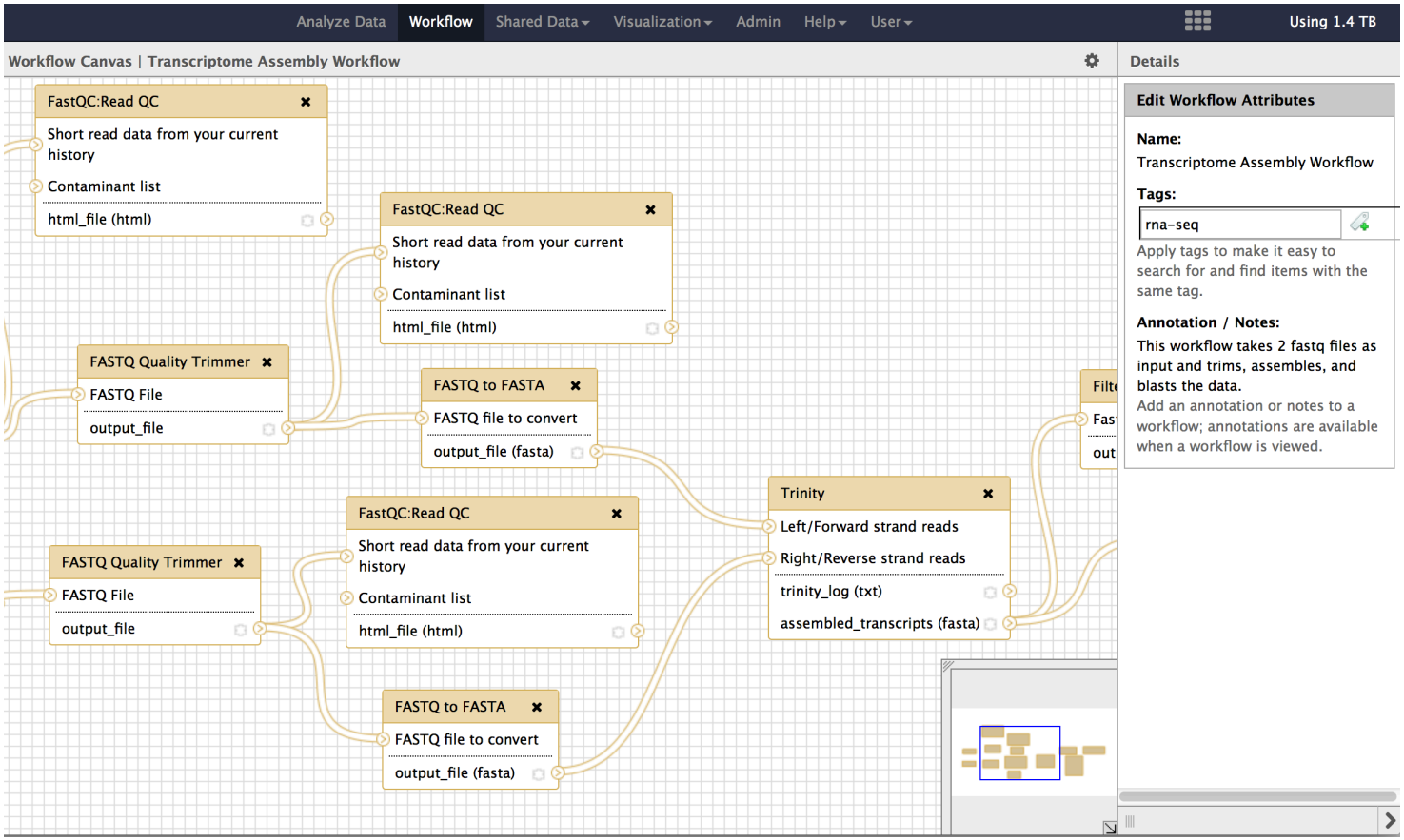
Adding more information

Renaming the history items makes a history much cleaner to follow – tracking back the datasets by number is not ideal

Galaxy allows annotation and note-taking per dataset in the history



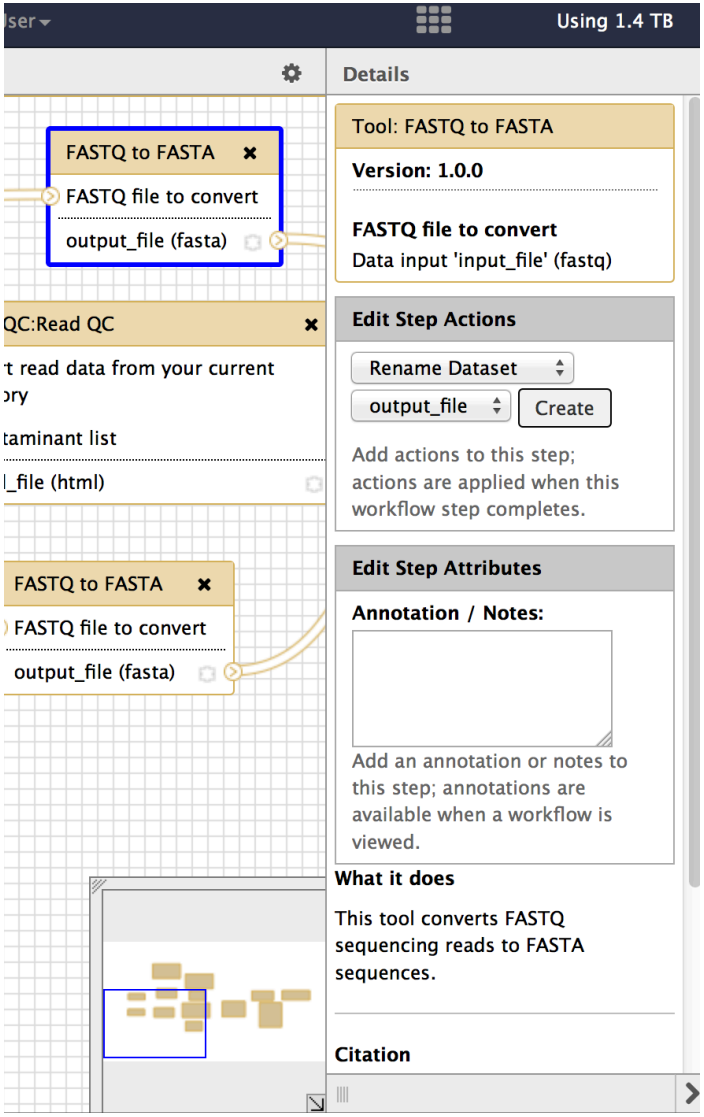
Adding more information



Notes can be added to workflows as well as histories

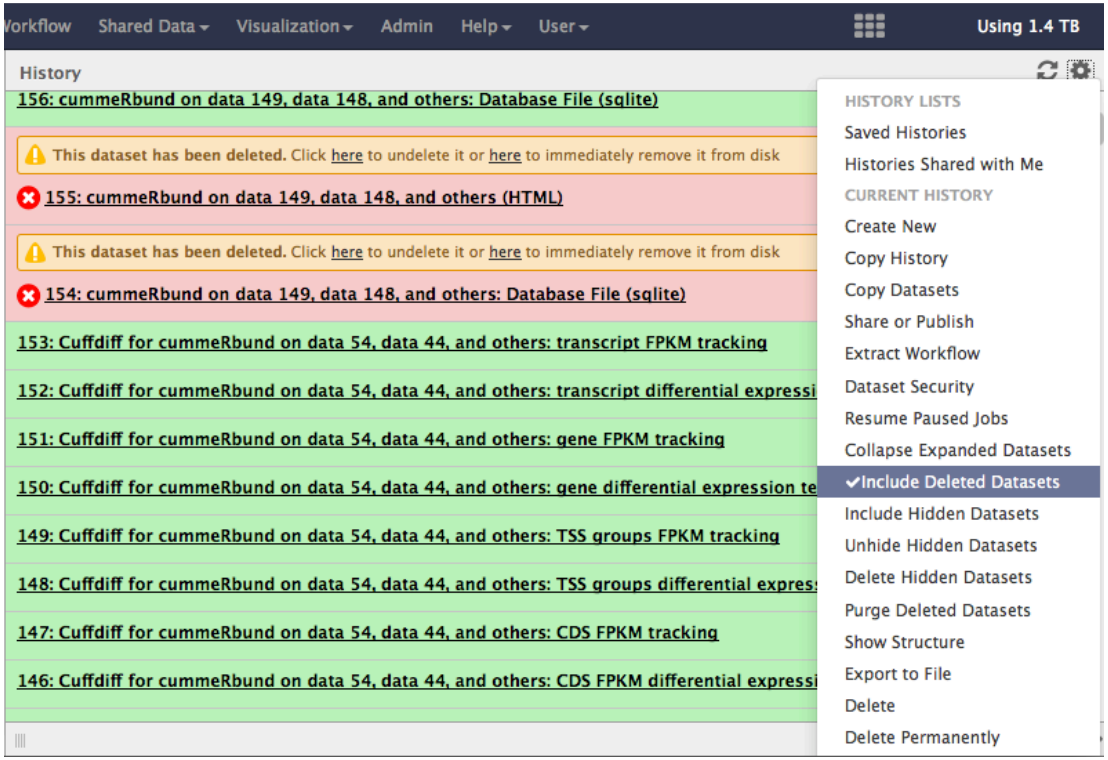
Adding more information

Individual steps can be renamed at runtime, as well.

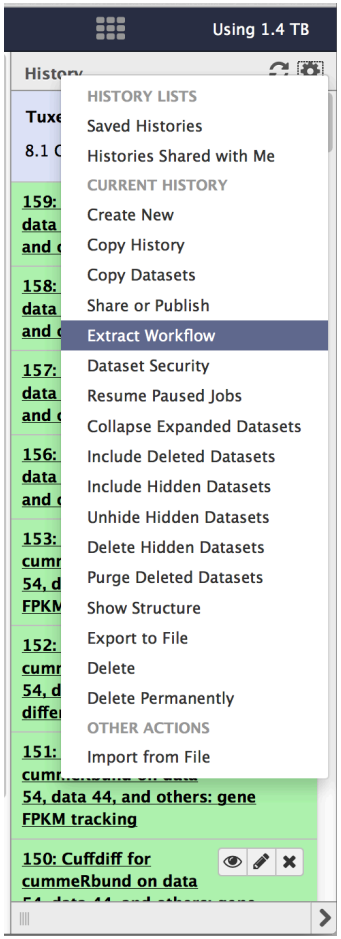


Backtracking

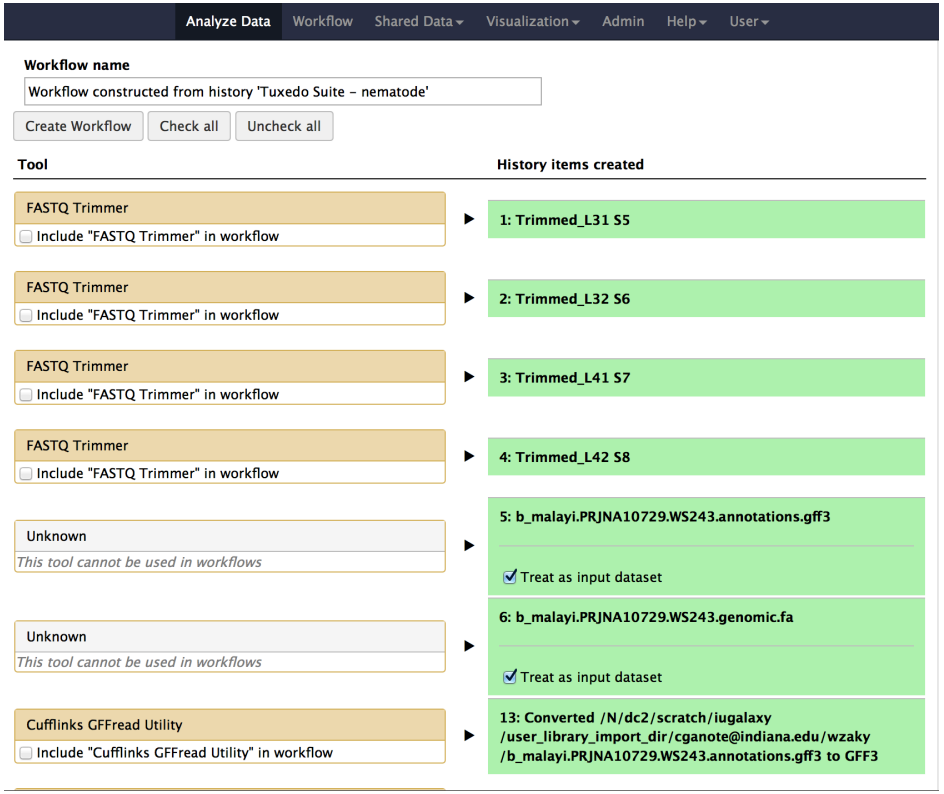
Items hidden/ deleted from the history are not gone forever – encouraging pruning of histories to retain the current workflow



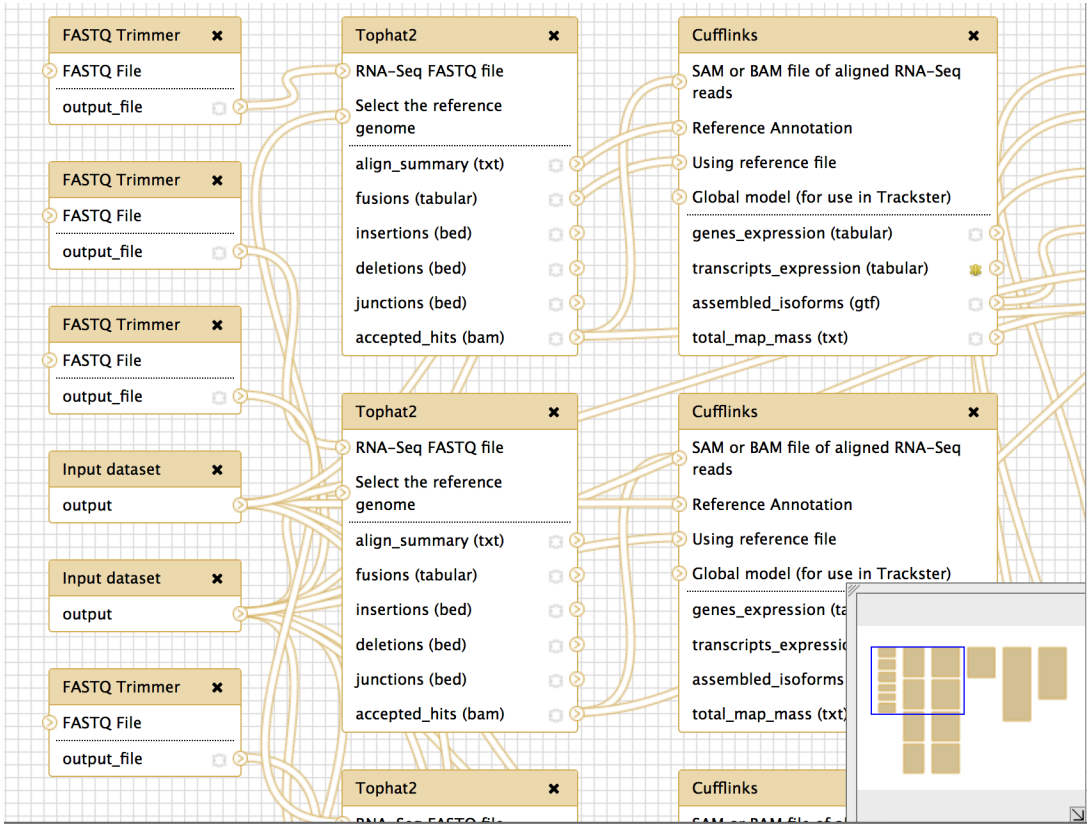
Visualizing Histories



Creating a workflow from a history is one way to visualize the pipeline, and also to ensure that the pipeline works as intended



Visualizing Histories



Pipelines can be quite complex – but if you are good at reading spaghetti, then this is the method for you

Provenance – how much is too much?

It is good to know and record, for example, which release of NR you ran your blast job on

Keep a copy of every release for sake of reproducing blast results?

We keep 121GB of Blast database available on DC2 and update monthly

What about obsolete software?

Different results may be derived from different machines – decommissioned machines are long gone

Planetary alignments?

End Goals

Galaxy workflows and raw inputs are enough to publish a perfectly reproducible experiment

The same results 1, 5 and 10 years from now

We are not quite there yet!

Fín

Thanks for watching!

Questions and comments:

Email help@ncgas.org